Wherefore, what is claimed is:

1. A computer-implemented process for detecting speech, comprising the process actions of:

inputting associated audio and video training data containing a person's face that is periodically speaking; and

using said audio and video signals to train a time delay neural network to determine when a person is speaking, wherein said training comprises the following process actions:

computing audio features from said audio training data wherein said audio feature is the energy over an audio frame;

computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and

correlating said audio features and video features to determine when a person is speaking.

2. The computer-implemented process of Claim 1 further comprising the process action of preprocessing the audio and video signals prior to using said audio and video signals to train a Time Delay Neural Network.

3. The computer-implemented process of Claim 2 wherein said process action of preprocessing the audio and video signals comprises the process actions of:

reducing the noise of the audio signals;

segmenting the audio data signals;

segmenting the video data signals;

extracting audio features; and

extracting video features.

4. The computer-implemented process of Claim 3 wherein the process action of segmenting the audio data signal comprises the process action of segmenting the audio data to determine regions of speech and non-speech.

5. The computer-implemented process of Claim 3 wherein the process action of segmenting the video data signal comprises the process action of segmenting the video data to determine at least one face and a mouth region within said determined faces.

6. The computer-implemented process of Claim 1 wherein the process action of computing video features from said video training signals comprises the process actions of:

using a face detector to locate a face in said video training signals;

using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image;

stabilizing the mouth image to remove any translational motion of the mouth caused by head movement;

using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed; and

designating values of mouth openness wherein the values range from -1 for the mouth being closed, to +1 for the mouth being open.

7. The computer-implemented process of Claim 6 wherein the process action of stabilizing the mouth image comprises the process action of using normalized cross correlation to remove any of said translational movement.

8. The computer-implemented process of Claim 6 wherein the process action of using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image comprises the process actions of:

using a generic model of a head, designating the mouth region to be centered at 0.7 of the head height from the top of the head model;

designating the width of the mouth as one quarter of the height of the head; and

designating the height of the mouth as 1/10th of the height of the head model.

9. The computer-process of Claim 1 further comprising the process actions of:

inputting an associated audio and video sequence of a person periodically speaking;

using said trained Time Delay Neural Network to determine when in said audio and video sequence said person is speaking.

10. The computer-implemented process of Claim 9 further comprising the process action of preprocessing the associated audio and video sequence prior to using said trained Time Delay Neural Network to determine if a person is speaking.

11. The computer-implemented process of Claim 11 wherein said process action of preprocessing the audio and video sequence comprises the process actions of:

reducing the noise of the audio in said sequence;

segmenting the audio data in said sequence;

segmenting the video signals in said sequence;

extracting audio features from said sequence; and

extracting video features from said sequence.

12. A computer-readable memory containing a computer program that is executable by a computer to perform the process recited in Claim 9.

13. A computer-readable medium having computer-executable instructions for use in detecting when a person in a synchronized audio video clip is speaking, said computer executable instructions comprising:

inputting one or more captured video and synchronized audio clips,

segmenting said audio and video clips to remove portions of said video and synchronized audio clips not needed in determining if a speaker in the captured video and synchronized audio clips is speaking;

extracting audio and video features in said captured video and synchronized audio clips to be used in determining if a speaker in the captured; and wherein an audio feature is the energy over an audio frame and wherein said video feature is the openness of a person's mouth;

training a Time Delay Neural Network to determine when a person is speaking using said extracted audio and video features.

15. The computer-readable medium of Claim 14 wherein the instruction for training a Time Delay Neural Network further comprises a sub-instruction for correlating said audio features and video features to determine when a person is speaking.

15. The computer-readable medium of Claim 13, further comprising instructions for:

inputting a captured video and synchronized audio clip for which it is desired to detect a person speaking;

using said trained Time Delay Neural Network to determine when a person is speaking in the captured video and synchronized audio clip for which it is desired to detect a person speaking by using said extracted audio and video features.

16. The computer-readable medium of Claim 13 further comprising an instruction for reducing noise in said audio video clips prior to said instruction for segmenting said audio and video clips.

17. The computer-readable medium of Claim 13 wherein the process action of extracting audio and video features in said captured video and comprises sub-instructions for extracting video features comprising:

using a face detector to locate a face in said video training signals;

using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image;

stabilizing the mouth image to remove any translational motion of the mouth caused by head movement;

using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed; and designating values of mouth openness wherein the values range from -1 for the mouth being closed, to +1 for the mouth being open.

18. The computer-readable medium of Claim 17 wherein said sub-instruction for stabilizing the mouth image to remove any translational motion of the mouth caused by head movement employs normalized cross correlation.

19. The computer-readable medium of Claim 16 wherein said sub-instruction for using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image, comprises sub-instructions for:

using a generic model of a head, designating the mouth region to be centered at a given distance of the head height from the top of the head model;

designating the width of the mouth as a percentage of the height of the head; and

designating the height of the mouth as a percentage of the height of the head model.

20. A system for detecting a speaker in a video segment that is synchronized with associated audio, the system comprising

a general purpose computing device; and

a computer program comprising program modules executable by the computing device, wherein the computing device is directed by the program modules of the computer program to, input one or more captured video and synchronized audio segments,

segment said audio and video segments to remove portions of said video and synchronized audio segments not needed in determining if a speaker in the captured video and synchronized audio segments is speaking;

extract audio and video features in said captured video and synchronized audio segments to be used in determining if a speaker in the captured video and synchronized audio segments is speaking, wherein said audio feature is the energy over an audio frame and said video feature is the openness of a person's mouth in said video and synchronized audio segments;

train a Time Delay Neural Network to determine when a person is speaking using said extracted audio and video features.

input a captured video and synchronized audio clip for which it is desired to detect a person speaking; and

use said trained Time Delay Neural Network to determine when a person is speaking in the captured video and synchronized audio segments for which it is desired to detect a person speaking.

21. The system of Claim 20 wherein the module for using said trained Time Delay Neural Network comprises a sub-module that outputs a 1 when a person is talking for each frame in said captured video and synchronized audio segments for which it is desired to detect a person speaking, and outputs a 0 when no person is talking.

22. The system of Claim 20 wherein said Time Delay Neural Network comprises:

an input layer;

two hidden layers; and

one output, wherein said output is set to 0 when no person in the video and synchronized audio segment is speaking; and wherein said output is set to 1 when a person in the video and synchronized audio segment is speaking.

24. The system of Claim 20 wherein the module for extracting audio and video features comprises sub-modules to extract the video features comprising:

using a face detector to locate a face in said video training signals;

using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image;

stabilizing the mouth image to remove any translational motion of the mouth caused by head movement; and

using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed.

25. A computer-implemented process for detecting speech in an audio-visual sequence wherein more than one person is speaking at a time, comprising the process actions of:

inputting associated audio and video training data containing more than one person's face wherein each person is periodically speaking at the same time as the other person or persons; and

using said audio and video signals to train a time delay neural network to determine which person is speaking at a given time, wherein said training comprises the following process actions:

computing audio features from said audio training data wherein said audio feature is the energy over an audio frame;

computing video features from said video training signals to determine whether a given person's mouth is open or closed; and

correlating said audio features and video features to determine when a given person is speaking.

26. The computer-implemented process of Claim 25 wherein the process action of computing video features from said video training signals comprises the process actions of:

using a face detector to locate each face in said video training signals;

using a microphone array to beam form on each face detected thereby filtering out sound not coming from the direction of the speaker to create beam formed audio training data;

using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image;

stabilizing the mouth image to remove any translational motion of the mouth caused by head movement;

using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed; and

designating values of mouth openness wherein the values range from -1 for the mouth being closed, to +1 for the mouth being open.

27. The computer-implemented process of Claim 26 wherein said audio feature is computed using said beam formed audio training data.

28. The computer-implemented process of Claim 27 further comprising the process actions of:

inputting an associated audio and video sequence of more than one person periodically speaking;

using said trained Time Delay Neural Network to determine when in said audio and video sequence each person is speaking.

29. A computer-implemented process for detecting speech, comprising the process actions of:

inputting associated audio and video training data containing a person's face that is periodically speaking; and using said audio and video signals to train a statistical learning engine to determine when a person is speaking, wherein said training comprises the following process actions:

computing audio features from said audio training;

computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and

correlating said audio features and video features to determine when a person is speaking.

30. The computer-implemented process of Claim 29 wherein said audio feature is the acoustical energy over an audio frame.

31. The computer-implemented process of Claim 29 wherein said audio feature is defined by Mel cepstrum coefficients.

32. The computer-implemented process of Claim 29 wherein said statistical learning engine is a Time Delay Neural Network.

33. The computer-implemented process of Claim 29 wherein said statistical learning engine is a Support Vector Machine.